Analyzing the Attributes of Wine Enthusiast's Wine Reviews

Final Project for NYU Data Bootcamp Class, Spring 2019

By Bethel Hailemichael and Carmina Magnani

Project Description: Our project focuses on analyzing the different attributes of the wines reviewed by the Wine Enthusiast magazine. We focus on analyzing how the different countries of origin, pricing, grape varieties, and descriptions compare to the scores given to the wines. We also test our hypothesis about the length of a wine's description and its possible correlation to the type of scoring it receives.

In [1]:
```
from IPython.display import Image, display
display(Image('https://www.iacac.org/wp-content/uploads/wine-tasting-1500x846.jpg', width=1900, unconfined=True))
```

Part 1: Preparing the Data for Analysis

Source: Our data comes from the following kaggle dataset (https://www.kaggle.com/zynicide/wine-reviews), which is based on data compiled from WineEnthusiast issues from June 15th, 2017 and November 22nd, 2017.

Importing Packages

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib as mpl
import seaborn as sns
sns.set(style='white', context='notebook', palette='deep')
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

In [3]:
```
df = pd.read_csv('winemag-data_first150k.csv')
```

In [4]:
```
df
```

Out[4]:

	Unnamed: 0	country	description	designation	points	price	province	region_1
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley
1	1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro
2	2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley
3	3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley
4	4	France	This is the top wine from La Bégude, named aft...	La Brûlade	95	66.0	Provence	Bandol
5	5	Spain	Deep, dense and pure from the opening bell, th...	Numanthia	95	73.0	Northern Spain	Toro
6	6	Spain	Slightly gritty black-fruit aromas include a s...	San Román	95	65.0	Northern Spain	Toro
7	7	Spain	Lush cedary black-fruit aromas are luxe and of...	Carodorum Único Crianza	95	110.0	Northern Spain	Toro
8	8	US	This re-named vineyard was formerly bottled as...	Silice	95	65.0	Oregon	Chehalem Mountains
9	9	US	The producer sources from two blocks of the vi...	Gap's Crown Vineyard	95	60.0	California	Sonoma Coast
10	10	Italy	Elegance, complexity and structure come togeth...	Ronco della Chiesa	95	80.0	Northeastern Italy	Collio

	Unnamed: 0	country	description	designation	points	price	province	region_1
11	11	US	From 18-year-old vines, this supple well-balan...	Estate Vineyard Wadensvil Block	95	48.0	Oregon	Ribbon Ridge
12	12	US	A standout even in this terrific lineup of 201...	Weber Vineyard	95	48.0	Oregon	Dundee Hills
13	13	France	This wine is in peak condition. The tannins an...	Château Montus Prestige	95	90.0	Southwest France	Madiran
14	14	US	With its sophisticated mix of mineral, acid an...	Grace Vineyard	95	185.0	Oregon	Dundee Hills
15	15	US	First made in 2006, this succulent luscious Ch...	Sigrid	95	90.0	Oregon	Willamette Valley
16	16	US	This blockbuster, powerhouse of a wine suggest...	Rainin Vineyard	95	325.0	California	Diamond Mountain District
17	17	Spain	Nicely oaked blackberry, licorice, vanilla and...	6 Años Reserva Premium	95	80.0	Northern Spain	Ribera del Duero
18	18	France	Coming from a seven-acre vineyard named after ...	Le Pigeonnier	95	290.0	Southwest France	Cahors
19	19	US	This fresh and lively medium-bodied wine is be...	Gap's Crown Vineyard	95	75.0	California	Sonoma Coast
20	20	US	Heitz has made this stellar rosé from the rare...	Grignolino	95	24.0	California	Napa Valley
21	21	Spain	Alluring, complex and powerful aromas of grill...	Prado Enea Gran Reserva	95	79.0	Northern Spain	Rioja

	Unnamed: 0	country	description	designation	points	price	province	region_1
22	22	Spain	Tarry blackberry and cheesy oak aromas are app...	Termanthia	95	220.0	Northern Spain	Toro
23	23	US	The apogee of this ambitious winery's white wi...	Giallo Solare	95	60.0	California	Edna Valley
24	24	US	San Jose-based producer Adam Comartin heads 1,...	R-Bar-R Ranch	95	45.0	California	Santa Cruz Mountains
25	25	New Zealand	Yields were down in 2015, but intensity is up,...	Maté's Vineyard	94	57.0	Kumeu	NaN
26	26	US	Bergström has made a Shea designate since 2003...	Shea Vineyard	94	62.0	Oregon	Willamette Valley
27	27	US	Focused and dense, this intense wine captures ...	Abetina	94	105.0	Oregon	Willamette Valley
28	28	US	Cranberry, baked rhubarb, anise and crushed sl...	Garys' Vineyard	94	60.0	California	Santa Lucia Highlands
29	29	US	This standout Rocks District wine brings earth...	The Funk Estate	94	60.0	Washington	Walla Walla Valley (WA)
...	...	...	...	...	...	...	...	...
150900	150900	Chile	Aromas of freshly cut lumber, complete with so...	Prima Reserva	81	13.0	Maipo Valley	NaN
150901	150901	Chile	Lavishly oaked, the fruit here struggles to ma...	Reserva	81	12.0	Maipo Valley	NaN
150902	150902	Chile	This medium weight Chardonnay offered aromas o...	Estate Bottled	81	10.0	Maipo Valley	NaN

	Unnamed: 0	country	description	designation	points	price	province	region_1
150903	150903	Chile	Very light berry and mint aromas open this aus...	120	81	7.0	Rapel Valley	NaN
150904	150904	Chile	A lot of Chilean Cabernets seem to have a dist...	NaN	81	10.0	Maipo Valley	NaN
150905	150905	Chile	There's not much point in making a reserve-sty...	Prima Reserva	80	13.0	Maipo Valley	NaN
150906	150906	France	This lovely wine, a Monopole, is already showi...	Clos des Reas	93	65.0	Burgundy	Vosne-Romanée
150907	150907	France	Rion holds back on the new oak, letting the pu...	Les Beaux-Monts	92	52.0	Burgundy	Vosne-Romanée
150908	150908	France	Another premier cru from Michel Gros, this one...	Aux Brulees	90	65.0	Burgundy	Vosne-Romanée
150909	150909	France	This is a lovely, fragrant Burgundy, with a sm...	Clos dea Argillieres	89	52.0	Burgundy	Nuits-St.-Georges
150910	150910	France	Scents of graham cracker and malted milk choco...	NaN	89	38.0	Burgundy	Chambolle-Musigny
150911	150911	France	This needs a good bit of breathing time, then ...	Les Chaliots	87	37.0	Burgundy	Nuits-St.-Georges
150912	150912	France	The nose is dominated by the attractive scents...	Les Charmes	87	65.0	Burgundy	Chambolle-Musigny
150913	150913	France	Inky and rustic, yet in a refined manner. This...	NaN	94	30.0	Rhône Valley	Châteauneuf-du-Pape

	Unnamed: 0	country	description	designation	points	price	province	region_1
150914	150914	US	Old-gold in color, and thick and syrupy. The a...	Late Harvest Cluster Select	94	25.0	California	Anderson Valley
150915	150915	US	Decades ago, Beringer's then-winemaker Myron N...	Nightingale	93	30.0	California	North Coast
150916	150916	US	An impressive wine that presents a full bouque...	J. Schram	93	65.0	California	Napa Valley
150917	150917	France	Light and elegant, this spicy, lively wine is ...	Brut Mosaïque	92	30.0	Champagne	Champagne
150918	150918	France	Jacquart makes a full-bodied, ripe style of Ch...	Cuvée Mosaïque	92	38.0	Champagne	Champagne
150919	150919	France	This classy example opens with a very floral n...	Cuvée President	91	37.0	Champagne	Champagne
150920	150920	Italy	Rich and mature aromas of smoke, earth and her...	Brut Riserva	91	19.0	Northeastern Italy	Trento
150921	150921	France	Shows some older notes: a bouquet of toasted w...	Blanc de Blancs Brut Mosaïque	91	38.0	Champagne	Champagne
150922	150922	Italy	Made by 30-ish Roberta Borghese high above Man...	Superiore	91	NaN	Northeastern Italy	Colli Orientali del Friuli
150923	150923	France	Rich and toasty, with tiny bubbles. The bouque...	Demi-Sec	91	30.0	Champagne	Champagne
150924	150924	France	Really fine for a low-acid vintage, there's an...	Diamant Bleu	91	70.0	Champagne	Champagne
150925	150925	Italy	Many people feel Fiano represents southern Ita...	NaN	91	20.0	Southern Italy	Fiano di Avellino

	Unnamed: 0	country	description	designation	points	price	province	region_1
150926	150926	France	Offers an intriguing nose with ginger, lime an...	Cuvée Prestige	91	27.0	Champagne	Champagne
150927	150927	Italy	This classic example comes from a cru vineyard...	Terre di Dora	91	20.0	Southern Italy	Fiano di Avellino
150928	150928	France	A perfect salmon shade, with scents of peaches...	Grand Brut Rosé	90	52.0	Champagne	Champagne
150929	150929	Italy	More Pinot Grigios should taste like this. A r...	NaN	90	15.0	Northeastern Italy	Alto Adige

150930 rows × 11 columns

Cleaning the Data

Removing NaN values

In [5]:
```
df.dropna(axis = 0, how ='any')
```

Out[5]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	
2	2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley	
3	3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley	Wil
8	8	US	This re-named vineyard was formerly bottled as...	Silice	95	65.0	Oregon	Chehalem Mountains	Wil
9	9	US	The producer sources from two blocks of the vi...	Gap's Crown Vineyard	95	60.0	California	Sonoma Coast	
11	11	US	From 18-year-old vines, this supple well-balan...	Estate Vineyard Wadensvil Block	95	48.0	Oregon	Ribbon Ridge	Wil
12	12	US	A standout even in this terrific lineup of 201...	Weber Vineyard	95	48.0	Oregon	Dundee Hills	Wil
14	14	US	With its sophisticated mix of mineral, acid an...	Grace Vineyard	95	185.0	Oregon	Dundee Hills	Wil
15	15	US	First made in 2006, this succulent luscious Ch...	Sigrid	95	90.0	Oregon	Willamette Valley	Wil
16	16	US	This blockbuster, powerhouse of a wine suggest...	Rainin Vineyard	95	325.0	California	Diamond Mountain District	
19	19	US	This fresh and lively medium-bodied wine is be...	Gap's Crown Vineyard	95	75.0	California	Sonoma Coast	

	Unnamed: 0	country	description	designation	points	price	province	region_1	
20	20	US	Heitz has made this stellar rosé from the rare...	Grignolino	95	24.0	California	Napa Valley	
23	23	US	The apogee of this ambitious winery's white wi...	Giallo Solare	95	60.0	California	Edna Valley	
24	24	US	San Jose-based producer Adam Comartin heads 1,...	R-Bar-R Ranch	95	45.0	California	Santa Cruz Mountains	
27	27	US	Focused and dense, this intense wine captures ...	Abetina	94	105.0	Oregon	Willamette Valley	Wil
28	28	US	Cranberry, baked rhubarb, anise and crushed sl...	Garys' Vineyard	94	60.0	California	Santa Lucia Highlands	
29	29	US	This standout Rocks District wine brings earth...	The Funk Estate	94	60.0	Washington	Walla Walla Valley (WA)	Co
31	31	US	Steely and perfumed, this wine sees only 20% n...	Babushka	90	37.0	California	Russian River Valley	
34	34	US	The aromas entice with notes of wet stone, hon...	Conner Lee Vineyard	90	42.0	Washington	Columbia Valley (WA)	Co
47	47	US	Blended with 9% Malbec, 9% Cabernet Franc and ...	Estate Grown	90	60.0	California	Mount Veeder	
55	55	US	A blend of Cabernet from Grand Ciel (31%), Cie...	Four Flags	90	69.0	Washington	Red Mountain	Co
57	57	US	While exuberantly fruity, almost tropical on t...	Reserve	90	25.0	New York	Finger Lakes	
58	58	US	Cabernet makes up just over half of this blend...	Final Final	90	30.0	Washington	Columbia Valley (WA)	Co

	Unnamed: 0	country	description	designation	points	price	province	region_1	
65	65	US	Fresh boysenberries and a blueberry sorbet cha...	Estate Select	91	36.0	California	Santa Clara Valley	
67	67	US	From the producer's monumental Atlas Peak vine...	Animo	91	85.0	California	Napa Valley	
68	68	US	Big, bold, dark and chewy, this builds upon su...	Schindler Vineyard	91	50.0	Oregon	Eola-Amity Hills	Wil
69	69	US	A juiciness of cherry and vanilla spark the op...	Barrel Select	91	60.0	California	Rutherford	
70	70	US	Sweetened tannins highlight a depth of chocola...	District Collection	91	85.0	California	St. Helena	
75	75	US	An elegant blend from different estate vineyar...	Premier Cuvée	91	54.0	Oregon	Willamette Valley	Wil
77	77	US	Given four months in concrete egg, this is an ...	Juliana Vineyard	91	38.0	California	Napa Valley	
...	...	...	...	...	...	...	...	...	
150751	150751	US	Plummy, ripe, beginning to oxidize. Ready to d...	Gold Leaf Cuvee	82	18.0	California	Dry Creek Valley	
150753	150753	US	Simple and fruity and forward, though still un...	Sin Zin	82	18.0	California	Sonoma County	
150755	150755	US	Dark, brick colored wine, heavily oxidized, wi...	Black Zinfandel	82	24.0	California	California	Ca
150757	150757	US	Full blueberry and blackberry aromas can't ove...	Grandmére	81	25.0	California	Amador County	S

	Unnamed: 0	country	description	designation	points	price	province	region_1	
150758	150758	US	The more I taste Vidal, the more I find it an ...	Ice Wine	81	18.0	New York	Finger Lakes	
150762	150762	US	This wine is still young and precocious, but d...	Private Reserve	95	100.0	California	Napa Valley	
150764	150764	US	Specially selected lots account for this top-o...	Heritage Reserve	93	30.0	California	Carneros	N
150769	150769	US	Dark, oaky, and fine in every respect. Sonoma ...	Massara Vineyard	92	25.0	California	Sonoma Valley	
150774	150774	US	A very distinctive Chard—full of character and...	Shop Block Vineyard Dutton Ranch	91	30.0	California	Russian River Valley	
150787	150787	US	This wine become pretty pricey but that's beca...	Thomas Road Vineyard	90	50.0	California	Sonoma County	
150788	150788	US	This Chard has always been dependably rich, an...	Byron Vineyard	90	32.0	California	Santa Maria Valley	
150790	150790	US	Always a bit herbaceous, this year the underly...	Meritage	90	28.0	California	Sonoma County	
150792	150792	US	Strong , complex aromas, with notes of tropica...	Ruxton Vineyard	89	27.0	California	Russian River Valley	
150794	150794	US	Spicy black fruit, with bell pepper, black pep...	Chiles Mill Vineyard	89	22.0	California	Napa Valley	
150796	150796	US	Made largely of Cabernet Franc, this wine has ...	Fathom	89	30.0	California	Santa Barbara County	
150803	150803	US	Here's a wine with great bones: nicely structu...	Lavender Hill Vineyard	88	33.0	California	Napa Valley	

	Unnamed: 0	country	description	designation	points	price	province	region_1	
150807	150807	US	Ripe, tropical fruit aromas mark this spicy, w...	Oliver's Vineyard	88	20.0	California	Edna Valley	
150812	150812	US	Ripe, muscular wine, with a full throttle nose...	Reserve	88	25.0	California	Napa Valley	
150827	150827	US	Opens with blackberries and spices; fruity, fo...	Epoch II Millenium Cuvée	87	60.0	California	Dry Creek Valley	
150831	150831	US	A rich, creamy wine, loaded with personality. ...	Dutton Vineyard	87	22.0	California	Russian River Valley	
150836	150836	US	With this wine, Callaway takes a step away fro...	Coastal	86	10.0	California	California	Ca
150856	150856	US	White Merlot? Why not? It's actually a deep ro...	Forest Fire	84	8.0	California	California	Ca
150861	150861	US	From a dependable producer and a fine vintage ...	Reserve	84	16.0	California	Sonoma County	
150873	150873	US	With a color akin to the soft pink of ripe pea...	White	83	7.0	California	Amador County	S
150883	150883	US	A coppery colored, off-dry-to-frankly-sweet wi...	Reserve White	83	7.0	California	California	Ca
150889	150889	US	A bizarre style of wine. The aromas are Port-l...	Lafond Vineyard	82	35.0	California	Santa Ynez Valley	
150892	150892	US	A light, earthy wine, with violet, berry and t...	Coastal	82	10.0	California	California	Ca
150914	150914	US	Old-gold in color, and thick and syrupy. The a...	Late Harvest Cluster Select	94	25.0	California	Anderson Valley	Me

	Unnamed: 0	country	description	designation	points	price	province	region_1
150915	150915	US	Decades ago, Beringer's then-winemaker Myron N...	Nightingale	93	30.0	California	North Coast
150916	150916	US	An impressive wine that presents a full bouque...	J. Schram	93	65.0	California	Napa Valley

39241 rows × 11 columns

Removing "Unnamed" column to simplify display

In [6]:
```
df.drop(columns=['Unnamed: 0'])
```

Out[6]:

	country	description	designation	points	price	province	region_1	regio
0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	N
1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	I
2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley	Son
3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley	Willamette Va
4	France	This is the top wine from La Bégude, named aft...	La Brûlade	95	66.0	Provence	Bandol	I
5	Spain	Deep, dense and pure from the opening bell, th...	Numanthia	95	73.0	Northern Spain	Toro	I
6	Spain	Slightly gritty black-fruit aromas include a s...	San Román	95	65.0	Northern Spain	Toro	I
7	Spain	Lush cedary black-fruit aromas are luxe and of...	Carodorum Único Crianza	95	110.0	Northern Spain	Toro	I
8	US	This re-named vineyard was formerly bottled as...	Silice	95	65.0	Oregon	Chehalem Mountains	Willamette Va
9	US	The producer sources from two blocks of the vi...	Gap's Crown Vineyard	95	60.0	California	Sonoma Coast	Son
10	Italy	Elegance, complexity and structure come togeth...	Ronco della Chiesa	95	80.0	Northeastern Italy	Collio	I

	country	description	designation	points	price	province	region_1	regio
11	US	From 18-year-old vines, this supple well-balan...	Estate Vineyard Wadensvil Block	95	48.0	Oregon	Ribbon Ridge	Willamette Va
12	US	A standout even in this terrific lineup of 201...	Weber Vineyard	95	48.0	Oregon	Dundee Hills	Willamette Va
13	France	This wine is in peak condition. The tannins an...	Château Montus Prestige	95	90.0	Southwest France	Madiran	I
14	US	With its sophisticated mix of mineral, acid an...	Grace Vineyard	95	185.0	Oregon	Dundee Hills	Willamette Va
15	US	First made in 2006, this succulent luscious Ch...	Sigrid	95	90.0	Oregon	Willamette Valley	Willamette Va
16	US	This blockbuster, powerhouse of a wine suggest...	Rainin Vineyard	95	325.0	California	Diamond Mountain District	N
17	Spain	Nicely oaked blackberry, licorice, vanilla and...	6 Años Reserva Premium	95	80.0	Northern Spain	Ribera del Duero	I
18	France	Coming from a seven-acre vineyard named after ...	Le Pigeonnier	95	290.0	Southwest France	Cahors	I
19	US	This fresh and lively medium-bodied wine is be...	Gap's Crown Vineyard	95	75.0	California	Sonoma Coast	Son
20	US	Heitz has made this stellar rosé from the rare...	Grignolino	95	24.0	California	Napa Valley	N
21	Spain	Alluring, complex and powerful aromas of grill...	Prado Enea Gran Reserva	95	79.0	Northern Spain	Rioja	I
22	Spain	Tarry blackberry and cheesy oak aromas are app...	Termanthia	95	220.0	Northern Spain	Toro	I

	country	description	designation	points	price	province	region_1	regio
23	US	The apogee of this ambitious winery's white wi...	Giallo Solare	95	60.0	California	Edna Valley	Central C
24	US	San Jose-based producer Adam Comartin heads 1,...	R-Bar-R Ranch	95	45.0	California	Santa Cruz Mountains	Central C
25	New Zealand	Yields were down in 2015, but intensity is up,...	Maté's Vineyard	94	57.0	Kumeu	NaN	I
26	US	Bergström has made a Shea designate since 2003...	Shea Vineyard	94	62.0	Oregon	Willamette Valley	I
27	US	Focused and dense, this intense wine captures ...	Abetina	94	105.0	Oregon	Willamette Valley	Willamette Va
28	US	Cranberry, baked rhubarb, anise and crushed sl...	Garys' Vineyard	94	60.0	California	Santa Lucia Highlands	Central C
29	US	This standout Rocks District wine brings earth...	The Funk Estate	94	60.0	Washington	Walla Walla Valley (WA)	Columbia Va
...	...	...	...	...	...	...	...	
150900	Chile	Aromas of freshly cut lumber, complete with so...	Prima Reserva	81	13.0	Maipo Valley	NaN	I
150901	Chile	Lavishly oaked, the fruit here struggles to ma...	Reserva	81	12.0	Maipo Valley	NaN	I
150902	Chile	This medium weight Chardonnay offered aromas o...	Estate Bottled	81	10.0	Maipo Valley	NaN	I
150903	Chile	Very light berry and mint aromas open this aus...	120	81	7.0	Rapel Valley	NaN	I

	country	description	designation	points	price	province	region_1	regio
150904	Chile	A lot of Chilean Cabernets seem to have a dist...	NaN	81	10.0	Maipo Valley	NaN	I
150905	Chile	There's not much point in making a reserve-sty...	Prima Reserva	80	13.0	Maipo Valley	NaN	I
150906	France	This lovely wine, a Monopole, is already showi...	Clos des Reas	93	65.0	Burgundy	Vosne-Romanée	I
150907	France	Rion holds back on the new oak, letting the pu...	Les Beaux-Monts	92	52.0	Burgundy	Vosne-Romanée	I
150908	France	Another premier cru from Michel Gros, this one...	Aux Brulees	90	65.0	Burgundy	Vosne-Romanée	I
150909	France	This is a lovely, fragrant Burgundy, with a sm...	Clos dea Argillieres	89	52.0	Burgundy	Nuits-St.-Georges	I
150910	France	Scents of graham cracker and malted milk choco...	NaN	89	38.0	Burgundy	Chambolle-Musigny	I
150911	France	This needs a good bit of breathing time, then ...	Les Chaliots	87	37.0	Burgundy	Nuits-St.-Georges	I
150912	France	The nose is dominated by the attractive scents...	Les Charmes	87	65.0	Burgundy	Chambolle-Musigny	I
150913	France	Inky and rustic, yet in a refined manner. This...	NaN	94	30.0	Rhône Valley	Châteauneuf-du-Pape	I
150914	US	Old-gold in color, and thick and syrupy. The a...	Late Harvest Cluster Select	94	25.0	California	Anderson Valley	Mendocino/L Cour

	country	description	designation	points	price	province	region_1	regio
150915	US	Decades ago, Beringer's then-winemaker Myron N...	Nightingale	93	30.0	California	North Coast	North C
150916	US	An impressive wine that presents a full bouque...	J. Schram	93	65.0	California	Napa Valley	N
150917	France	Light and elegant, this spicy, lively wine is ...	Brut Mosaïque	92	30.0	Champagne	Champagne	I
150918	France	Jacquart makes a full-bodied, ripe style of Ch...	Cuvée Mosaïque	92	38.0	Champagne	Champagne	I
150919	France	This classy example opens with a very floral n...	Cuvée President	91	37.0	Champagne	Champagne	I
150920	Italy	Rich and mature aromas of smoke, earth and her...	Brut Riserva	91	19.0	Northeastern Italy	Trento	I
150921	France	Shows some older notes: a bouquet of toasted w...	Blanc de Blancs Brut Mosaïque	91	38.0	Champagne	Champagne	I
150922	Italy	Made by 30-ish Roberta Borghese high above Man...	Superiore	91	NaN	Northeastern Italy	Colli Orientali del Friuli	I
150923	France	Rich and toasty, with tiny bubbles. The bouque...	Demi-Sec	91	30.0	Champagne	Champagne	I
150924	France	Really fine for a low-acid vintage, there's an...	Diamant Bleu	91	70.0	Champagne	Champagne	I
150925	Italy	Many people feel Fiano represents southern Ita...	NaN	91	20.0	Southern Italy	Fiano di Avellino	I
150926	France	Offers an intriguing nose with ginger, lime an...	Cuvée Prestige	91	27.0	Champagne	Champagne	I

	country	description	designation	points	price	province	region_1	regio
150927	Italy	This classic example comes from a cru vineyard...	Terre di Dora	91	20.0	Southern Italy	Fiano di Avellino	I
150928	France	A perfect salmon shade, with scents of peaches...	Grand Brut Rosé	90	52.0	Champagne	Champagne	I
150929	Italy	More Pinot Grigios should taste like this. A r...	NaN	90	15.0	Northeastern Italy	Alto Adige	I

150930 rows × 10 columns

Part 2: Preliminary Visualization

First, let's take a look at the top 25 wines in the ranking. We'll sort by country and price too to get a sense of the range of complexity of the data.

In [7]:
```
df1 = df.sort_values(by=['points'], ascending = False)
df1 = df1[['points','country','price']]
df1 = df1.head(25)
df1
```

Out[7]:

	points	country	price
137099	100	US	200.0
19354	100	US	65.0
84035	100	Australia	300.0
84034	100	US	65.0
122767	100	US	100.0
2145	100	France	848.0
83536	100	France	1400.0
26296	100	France	1400.0
89399	100	US	200.0
24151	100	Italy	460.0
138867	100	Italy	210.0
41521	100	Italy	460.0
28954	100	Italy	195.0
119195	100	Australia	300.0
119194	100	US	65.0
92916	100	US	215.0
119521	100	Italy	460.0
19355	100	Australia	300.0
51886	100	France	1400.0
78004	100	Italy	195.0
98647	100	US	100.0
111087	100	Italy	210.0
114272	100	US	245.0
143522	100	US	245.0
26297	99	France	385.0

Per the results above, it seems that most of the wines (24 of them) share the highest score (100 points). Let's take a look now at the point distribution for all of the wines in the dataframe.

In [8]:
```
df.points.describe()
```

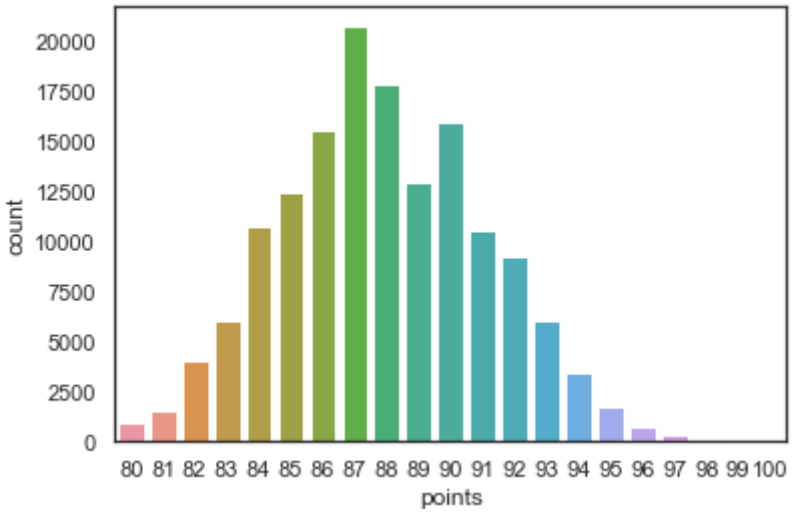
Out[8]:
```
count    150930.000000
mean         87.888418
std           3.222392
min          80.000000
25%          86.000000
50%          88.000000
75%          90.000000
max         100.000000
Name: points, dtype: float64
```

We can see that from the 150,930 wines in the list, the mean point score for all wines is 87.9 and the minimum score is 80. This demonstrates that the list compiles the highest ranked wines from all over the world.

Additionally, let's look at the points distribution in a graphic format

In [9]:
```
sns.countplot(df['points'])
```

Out[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x105918940>
```



Similarly, we can investigate if there is a wider variety in the distribution of the wines according to their price point.

In [10]:
```
df.price.describe()
```

Out[10]:
```
count    137235.000000
mean         33.131482
std          36.322536
min           4.000000
25%          16.000000
50%          24.000000
75%          40.000000
max        2300.000000
Name: price, dtype: float64
```

Here, we see that the average price point for the wines whose prices are recorded is $33.13, with the high-end price being \$2,300 and the low-end being $4.00

Next, let's take a look at the number of countries that are represented on the list.

In [11]:
```
countries_reviewed1 = df['country'].nunique()
countries_reviewed = "There are {} countries in this list".format(countr
ies_reviewed1)
print(countries_reviewed)
```

```
There are 48 countries in this list
```

Next, for each of the countries, let's sort them by the number of wines reviewed and also look at their lowest and highest priced wines.

In [12]:
```
countries_reviewed = df.groupby(['country']).price.agg([len, min, max])
countries_reviewed.sort_values(by='len')
```

Out[12]:

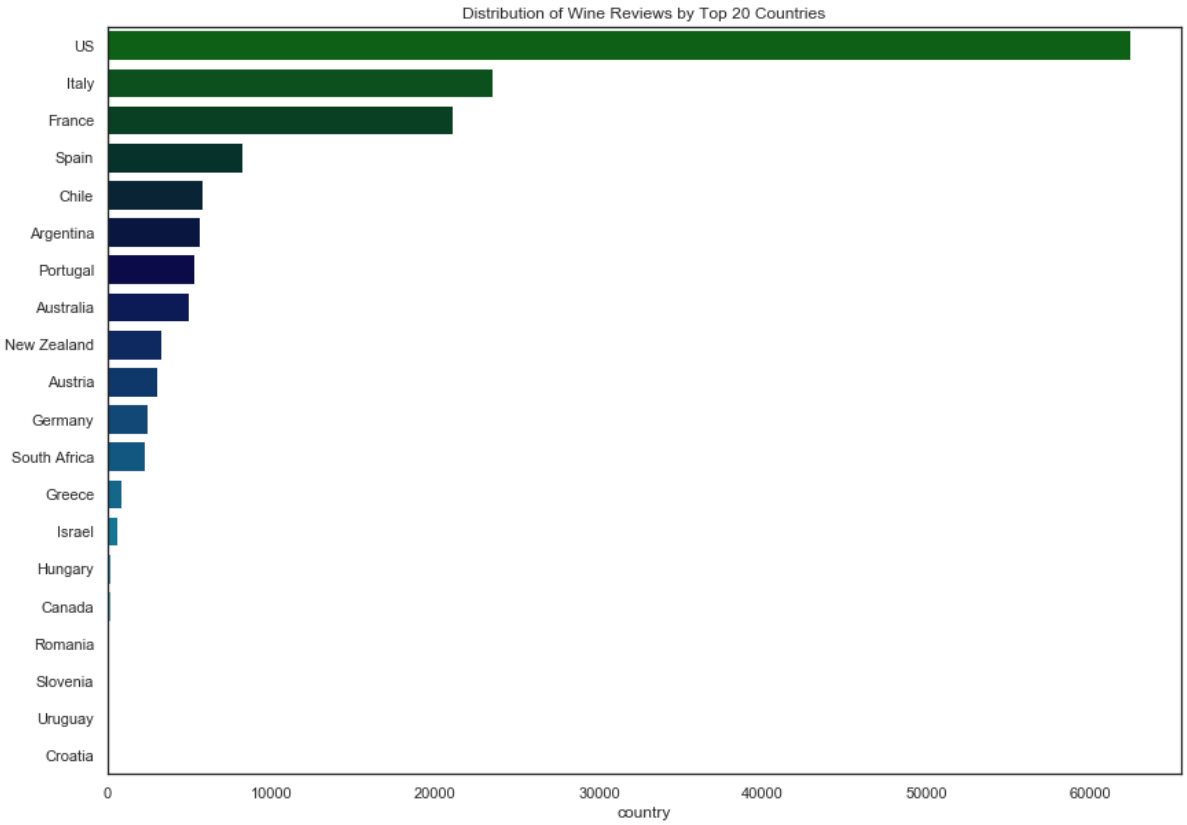
country	len	min	max
US-France	1.0	50.0	50.0
Albania	2.0	20.0	20.0
Tunisia	2.0	NaN	NaN
Montenegro	2.0	10.0	10.0
Japan	2.0	24.0	24.0
Slovakia	3.0	15.0	16.0
China	3.0	7.0	27.0
Egypt	3.0	NaN	NaN
Switzerland	4.0	19.0	38.0
South Korea	4.0	11.0	16.0
Bosnia and Herzegovina	4.0	12.0	13.0
Ukraine	5.0	13.0	13.0
Czech Republic	6.0	15.0	25.0
Lithuania	8.0	10.0	10.0
India	8.0	10.0	20.0
Luxembourg	9.0	36.0	50.0
England	9.0	38.0	75.0
Morocco	12.0	6.0	35.0
Serbia	14.0	15.0	42.0
Macedonia	16.0	12.0	25.0
Brazil	25.0	11.0	35.0
Cyprus	31.0	10.0	22.0
Lebanon	37.0	12.0	51.0
Georgia	43.0	9.0	40.0
Turkey	52.0	14.0	120.0
Mexico	63.0	12.0	108.0
Moldova	71.0	8.0	42.0
Bulgaria	77.0	7.0	28.0
Croatia	89.0	12.0	65.0
Uruguay	92.0	7.0	60.0
Slovenia	94.0	7.0	90.0
Romania	139.0	4.0	320.0
Canada	196.0	12.0	145.0

country	len	min	max
Hungary	231.0	7.0	764.0
Israel	630.0	8.0	150.0
Greece	884.0	7.0	120.0
South Africa	2258.0	5.0	145.0
Germany	2452.0	5.0	775.0
Austria	3057.0	8.0	1100.0
New Zealand	3320.0	7.0	125.0
Australia	4957.0	5.0	850.0
Portugal	5322.0	4.0	980.0
Argentina	5631.0	4.0	250.0
Chile	5816.0	5.0	400.0
Spain	8268.0	4.0	770.0
France	21098.0	5.0	2300.0
Italy	23478.0	5.0	900.0
US	62397.0	4.0	2013.0

As we can see, the country with the highest quantity of wines reviewed is the US. Below is a visualization of the countries and their wine review count.

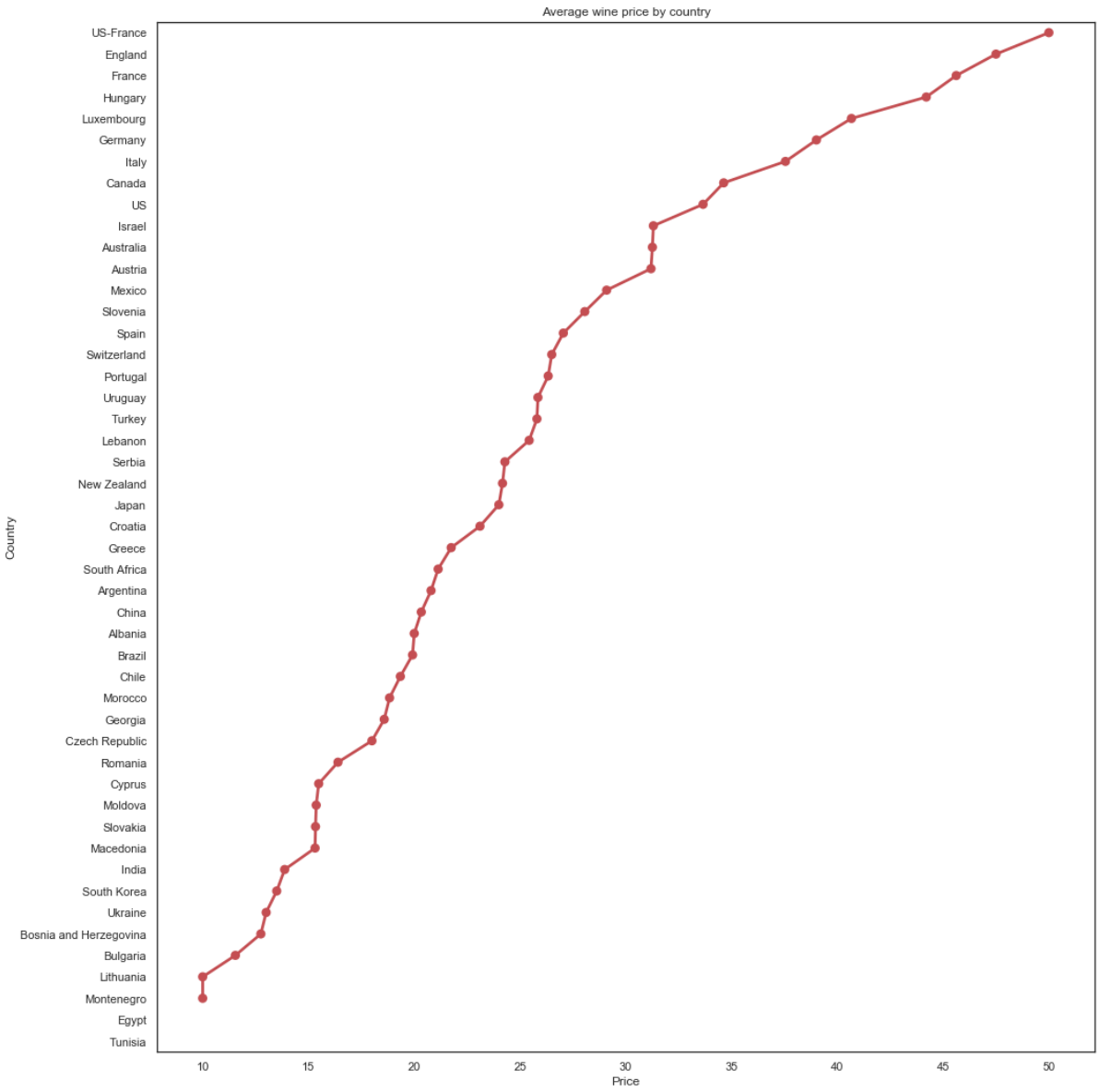
In [13]:
```
print('Number of country list in data:',df['country'].nunique())
plt.figure(figsize=(14,10))
cnt = df['country'].value_counts().to_frame()[0:20]
sns.barplot(x= cnt['country'], y =cnt.index, data=cnt, palette='ocean',o
rient='h')
plt.title('Distribution of Wine Reviews by Top 20 Countries');
```

```
Number of country list in data: 48
```



Lastly, we can take a look at the relationship between the countries listed and the average price of the wines reviewed.

In [14]:
```
cnt = df.groupby(['country',]).mean()['price'].sort_values(ascending=False).to_frame()

plt.figure(figsize=(16,18))
sns.pointplot(x = cnt['price'] ,y = cnt.index ,color='r',orient='h',markers='o')
plt.title('Average wine price by country')
plt.xlabel('Price')
plt.ylabel('Country');
```



Part 3: Testing if Description Length is a Predictor of Points Awarded

For this section, we'll evaluate whether a wine's description can be used to predict its score. First, let's start by importing a few packages that will help us visualize the data.

In [15]:
```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.metrics import classification_report, confusion_matrix, coh
en_kappa_score, roc_auc_score, roc_curve
from sklearn.feature_extraction import stop_words
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransf
ormer
from sklearn.decomposition import TruncatedSVD
from sklearn.feature_extraction import text
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.svm import LinearSVC
from sklearn.naive_bayes import MultinomialNB
from scipy import stats
from time import time
```

We will then start by simplifying the data. We will work with the columns that are relevant for our analysis and rename them.

In [16]:
```
df1 = df[['description', 'points']]
df1.rename(columns={'description':'Description',
                    'points':'Score'},
                    inplace=True)
```

Additionally, we want to isolate any records with empty descriptions, so we will add a column for the description length.

In [17]:
```
df1["Description_Length"] = [len(desc) for desc in df1['Description']]
```

Now, we will check if there are any missing values in our new column.

In [18]:
```
print("Number of missing values for the Score feature: ", len(df1[df1['S
core'].isnull()]))
print("Number of missing descriptions: ", len(df1[df1['Description_Lengt
h']==0]))
```

```
Number of missing values for the Score feature:  0
Number of missing descriptions:  0
```

As we can see, there are no missing descriptions in the list.

We will now proceed to test our hypothesis: whether the length of a wine's description is useful for predicting a wine's score.

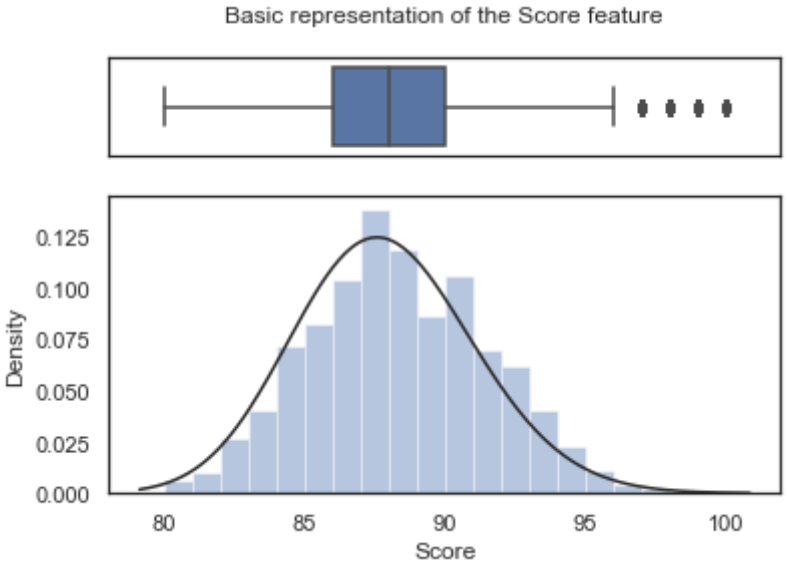
In [19]:
```
df1.describe()
```

Out[19]:

	Score	Description_Length
count	150930.000000	150930.000000
mean	87.888418	240.373948
std	3.222392	69.196308
min	80.000000	17.000000
25%	86.000000	193.000000
50%	88.000000	236.000000
75%	90.000000	282.000000
max	100.000000	829.000000

By looking at the statistics associated with the description, we can see that the average length for this field is 240 words. Let's drill further into the wine scores.

In [20]:
```
f, (ax_box, ax_hist) = plt.subplots(2, sharex=True,
                                    gridspec_kw={"height_ratios": (0.25,
0.75)})
sns.boxplot(df1["Score"], ax=ax_box).set_title("Basic representation of
 the Score feature\n")
sns.distplot(df1["Score"], ax=ax_hist, kde=False, fit=stats.gamma, bins=
20)
ax_box.set(xlabel='')
ax_hist.set(ylabel='Density')
plt.show()
```

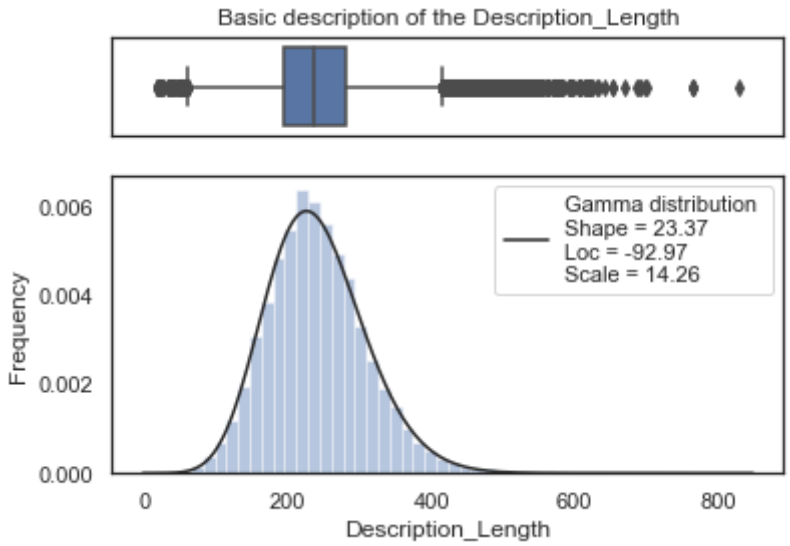


In [21]:
```
Q3 = np.quantile(df1['Score'], 0.75)
Q1 = np.quantile(df1['Score'], 0.25)
IQR = Q3 - Q1

outlier_score_threshold =  Q3 + 1.5 * IQR
outlier_number=len(df1[ df1['Score'] > outlier_score_threshold ])

print("Number of outliers:", outlier_number,
      "\nOutlier proportion:", round(outlier_number/len(df1['Score'])*10
0, 3),"%",
      "\nOutlier threshold score:", outlier_score_threshold,"/ 100")
```

```
Number of outliers: 570
Outlier proportion: 0.378 %
Outlier threshold score: 96.0 / 100
```

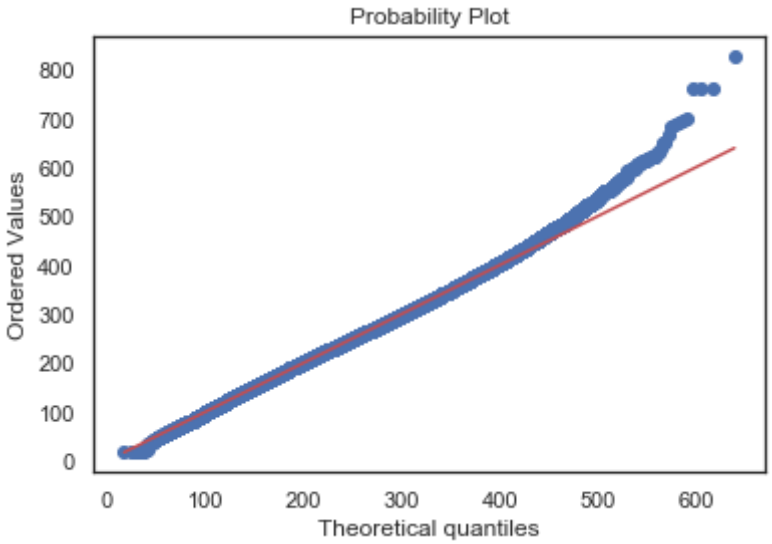
We can now observe that out of the 150k+ wines, only 570 of them stand out as outliers with a score that exceeds 96 points. This may be due to either tougher standards for the top-tier wines and/or scarcity in the amount of wines that meet this score.

In [22]:
```
f, (ax_box, ax_hist) = plt.subplots(2, sharex=True,
                                    gridspec_kw={"height_ratios": (0.25,
0.75)})
sns.boxplot(df1["Description_Length"],
            ax=ax_box).set_title("Basic description of the Description_L
ength")
sns.distplot(df1["Description_Length"],
             ax=ax_hist, kde=False, fit=stats.gamma, bins=50)
alpha, loc, beta = stats.gamma.fit(df1['Description_Length'])
plt.legend(['Gamma distribution \nShape = {0:.2f} \nLoc = {1:.2f}  \nSca
le = {2:.2f}'
            .format(alpha, loc, beta)],loc='best')
ax_box.set(xlabel='')
ax_hist.set(ylabel='Frequency')
plt.show()
```



Above we can see that the length of the descriptions is mostly concentrated around 100 and 400 words. Descriptions with 500+ words tend to be outliers

In [23]:
```
fig = plt.figure()
res = stats.probplot(df1['Description_Length'],
                     dist=stats.gamma(a= alpha, loc=loc, scale=beta), pl
ot=plt)
plt.show()
```



The plot above indicates that the deviation from the gamma distribution lies in descriptions with a length that exceeds 400 words. Now we will further segment the data into quartiles in order to observe the characteristics of the description field.

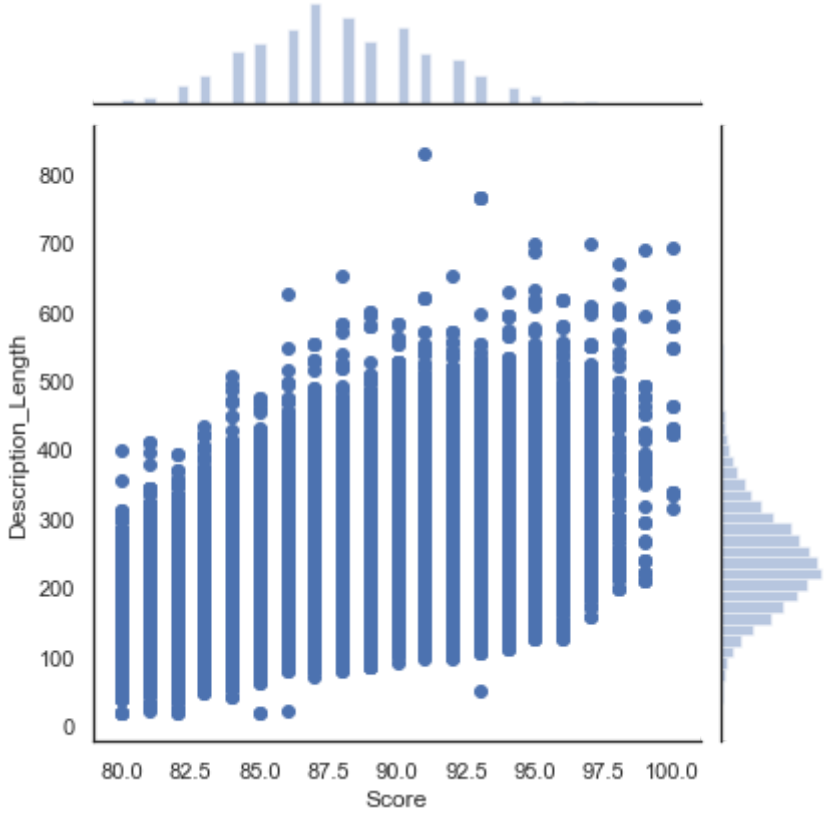
In [24]:
```
Q3 = np.quantile(df1['Description_Length'], 0.75)
Q1 = np.quantile(df1['Description_Length'], 0.25)
IQR = Q3 - Q1

outlier_score_threshold_high = Q3 + 1.5 * IQR
outlier_score_threshold_low = Q1 - 1.5 * IQR

outlier_number_total=len(df1[np.logical_or(df1['Description_Length'] >
                                           outlier_score_threshold_high,
                           df1['Description_Length'] < outlier_score_thres
hold_low)])

outlier_number_low = len(df1[df1['Description_Length'] < outlier_score_t
hreshold_low])
outlier_number_high = outlier_number_total - outlier_number_low

print("Number of outliers (high - low):", outlier_number_total,
      "(",outlier_number_high,"-",outlier_number_low,")",
      "\nOutlier proportion:",
      round(outlier_number_total/len(df1['Description_Length'])*100, 3),
"%",
      "\nOutlier threshold lengths (high - low):",
      outlier_score_threshold_high,"-",outlier_score_threshold_low)
```

```
Number of outliers (high - low): 2208 ( 2087 - 121 )
Outlier proportion: 1.463 %
Outlier threshold lengths (high - low): 415.5 - 59.5
```

As we can see, outliers make up about 1.463% of the data set.

We believe it would be interesting to test whether there is any correlation between the length of the description and the score given to a wine. Let's use a scatterplot to visualize this relationship.

In [25]:
```
sns.jointplot(x="Score", y="Description_Length", data=df1)
plt.show()

corr= np.corrcoef(df1["Score"], df1["Description_Length"])[0,1]
print("Correlation between Score and Description_Length:",round(corr,2))
```




```
Correlation between Score and Description_Length: 0.52
```

As we can see, the correlation is not quite strong (coefficient is below 0.7), but it is still positive. We cannot confirm there is a significant relationship between these two data points.

Part 4: Finding the Top 15 Most Frequently Used Words in Wine Descriptions

We will begin by separating the description from the scores.

In [26]:
```
corpus = df1["Description"].values
Y = df1["Score"].values
```

In [27]:
```
customStopWords = text.ENGLISH_STOP_WORDS.union(['wine', '2009', '2010',
'2011', '2012', '2013', '2014', '2015','2016', '2017', '2018',
                                                 '2019', '2020', '2021',
'2022','2023', '2024', '2025', '2030', '100', '10', '12',
                                                 '14', '15', '20', '25',
'30','40', '50', '60', '70', '90'])

CV = CountVectorizer(stop_words=customStopWords, max_features=1000, ngra
m_range=(1,2))
X = CV.fit_transform(corpus)

print("Number of entries (rows):", X.shape[0],\
      "\nNumber of features (columns):", X.shape[1])
```

```
Number of entries (rows): 150930
Number of features (columns): 1000
```

In [28]:
```
X_array = X.toarray()

inverted_dict = dict([[v,k] for k,v in CV.vocabulary_.items()])
final_dict = {}

for x in range(len(X_array[0,:])):
    final_dict[inverted_dict[x]]=np.sum(X_array[:,x])

print("15 most frequent words:",sorted(final_dict.items(),
                                        key = lambda kv:(kv[1], kv[0]), r
everse=True)[0:15])
```

```
15 most frequent words: [('flavors', 77992), ('fruit', 60474), ('finis
h', 37777), ('aromas', 35861), ('cherry', 32770), ('acidity', 32662),
('tannins', 32240), ('palate', 29404), ('ripe', 27096), ('black', 2659
1), ('dry', 24942), ('drink', 24146), ('spice', 23096), ('sweet', 2257
9), ('rich', 21433)]
```

We can now look at the list of 15 most frequently used words along with the number of instances these appear in the wine descriptions.

Part 5: Observations on Varieties

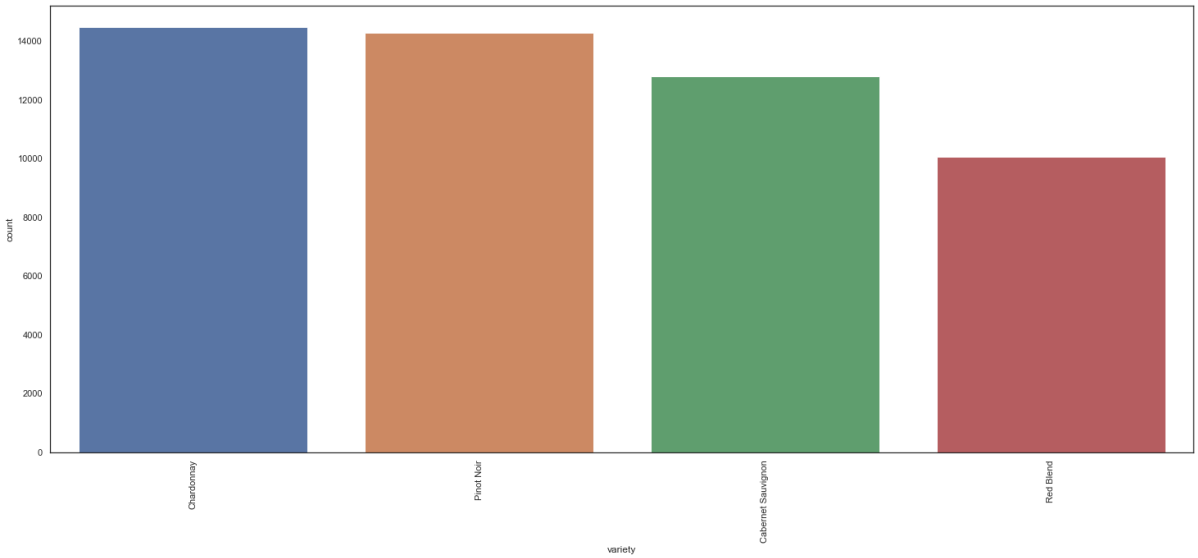
Lastly, let's take a look at the varieties of wine in the list.

In [29]:
```
df.variety.describe()
```

Out[29]:
```
count         150930
unique           632
top       Chardonnay
freq           14482
Name: variety, dtype: object
```

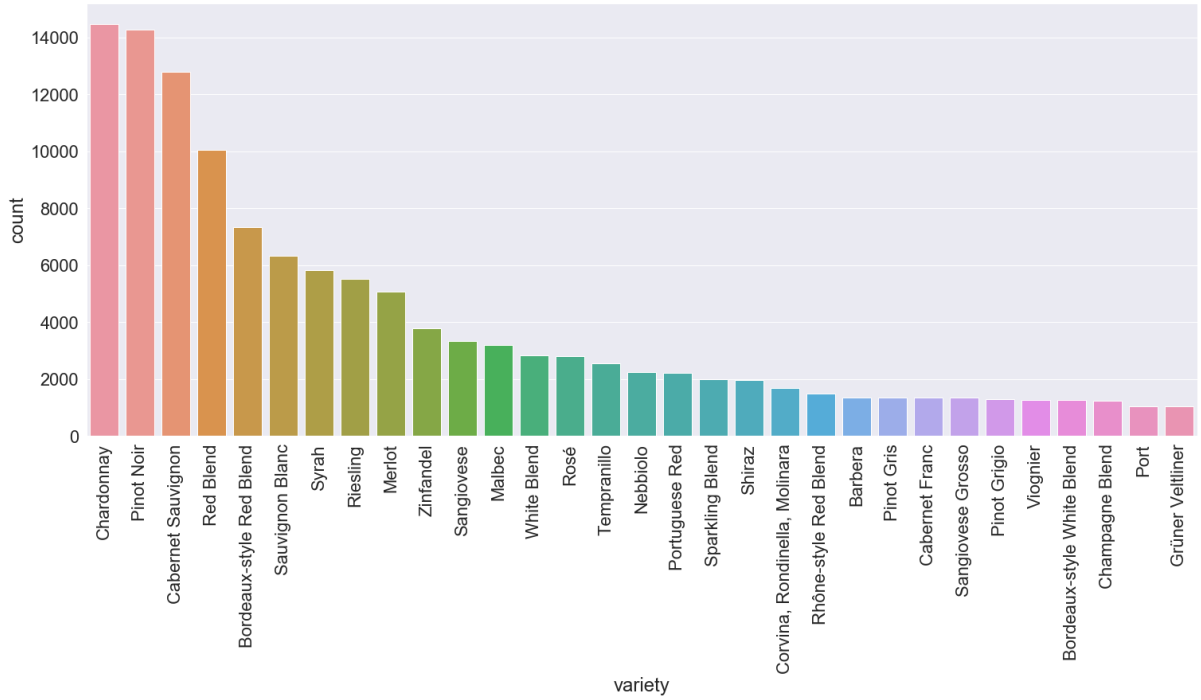
We can see that there are a total of 632 varieties, with Chardonnay being the most frequently listed.

In [30]:
```
variety_df = df.groupby('variety').filter(lambda x: len(x) > 10000)
varieties = variety_df['variety'].value_counts().index.tolist()
fig, ax = plt.subplots(figsize = (25, 10))
sns.countplot(x = variety_df['variety'], order = varieties, ax = ax)
sns.set(font_scale = 2)
plt.xticks(rotation = 90)
plt.show()
```



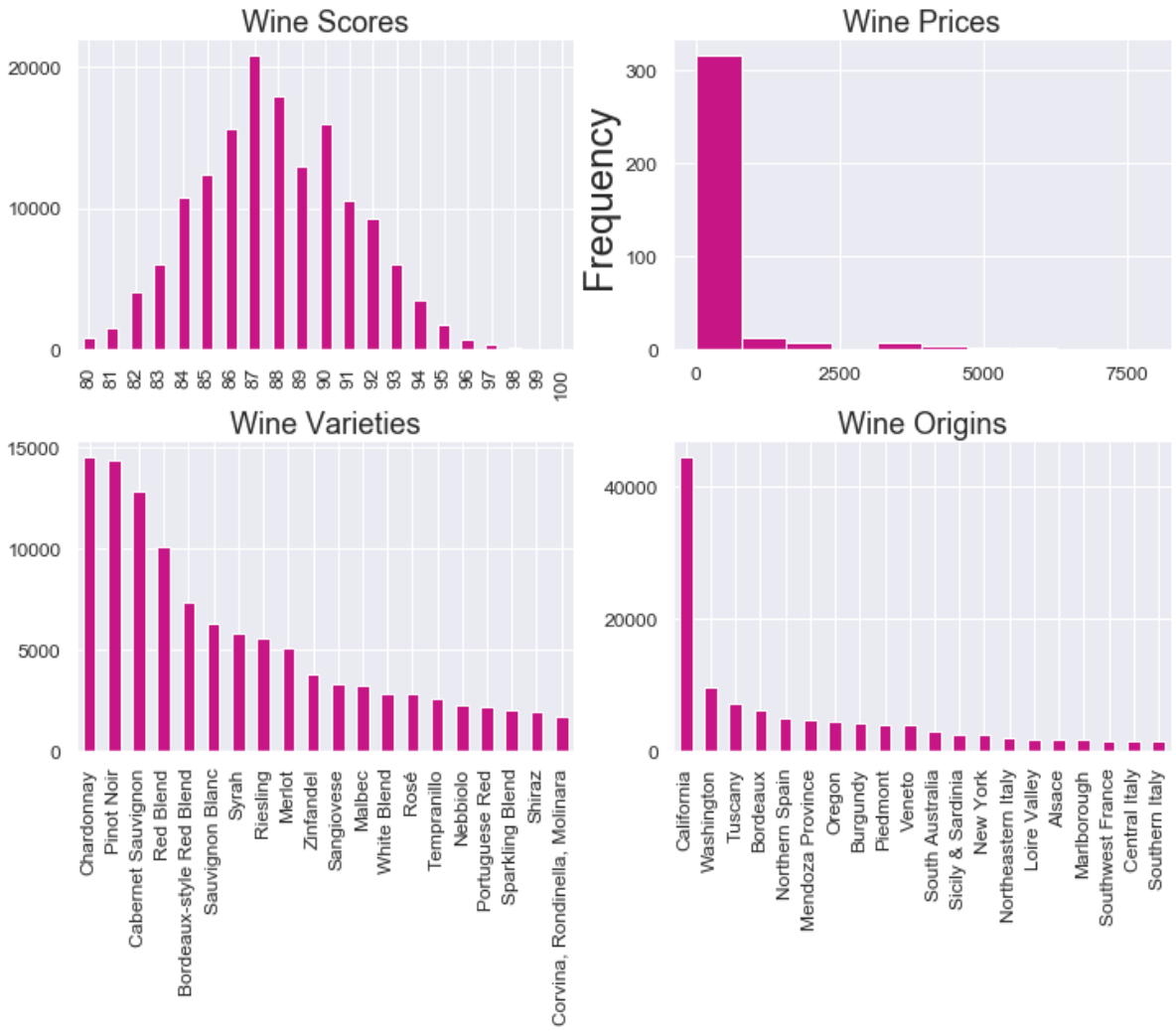
By narrowing down to varieties that are listed at least 10,000 times, we realize that there are four main varieties: Chardonnay, Pinot Noir, Cabernet Sauvignon, and Red Blend.

In [31]:
```
variety_df = df.groupby('variety').filter(lambda x: len(x) > 1000)
varieties = variety_df['variety'].value_counts().index.tolist()
fig, ax = plt.subplots(figsize = (25, 10))
sns.countplot(x = variety_df['variety'], order = varieties, ax = ax)
sns.set(font_scale = 2)
plt.xticks(rotation = 90)
plt.show()
```



Additionally, if we look at varieties that are listed at least 1,000 times, the data set demonstrates a lot more flavor options.

In [32]:
```
fig, axarr = plt.subplots(2, 2, figsize=(12, 8))

df['points'].value_counts().sort_index().plot.bar(
    ax=axarr[0][0], fontsize=12, color='mediumvioletred'
)
axarr[0][0].set_title("Wine Scores", fontsize=18)

df['variety'].value_counts().head(20).plot.bar(
    ax=axarr[1][0], fontsize=12, color='mediumvioletred'
)
axarr[1][0].set_title("Wine Varieties", fontsize=18)

df['province'].value_counts().head(20).plot.bar(
    ax=axarr[1][1], fontsize=12, color='mediumvioletred'
)
axarr[1][1].set_title("Wine Origins", fontsize=18)

df['price'].value_counts().plot.hist(
    ax=axarr[0][1], fontsize=12, color='mediumvioletred'
)
axarr[0][1].set_title("Wine Prices", fontsize=18)

plt.subplots_adjust(hspace=.3)

import seaborn as sns
sns.despine()
```



Lastly, in order to evaluate the "big picture" of our dataset, we have summarized a few of our findings in the graphs above.

We hope that our observations and analysis provide the readers (and wine enthusiasts!) with insight on the many ways that pricing, grape varieties, and country of origin relate to wine ratings.

Cheers!

Bethel Hailemichael and Carmina Magnani